

# Comparative analysis of three *Frankia* genomes: symbiotic interaction with actinorhizal host plants and genome plasticity.

Pascal Lapierre<sup>1</sup>, J. P. Gogarten<sup>1</sup>, Y. Huang<sup>1</sup>, J. Mastronunzio<sup>1</sup>, T. Rawnsley<sup>2</sup>, C. A. Bassi<sup>1</sup>, L. S. Tisa<sup>2</sup>, M. P. Francino<sup>4</sup>, Alla Lapidus<sup>4</sup>, P. Richardson<sup>4</sup>, P. Normand<sup>3</sup>, and D. R. Benson<sup>1</sup>

<sup>1</sup>Univ. Connecticut, Storrs, <sup>2</sup>Univ. New Hampshire, Durham, <sup>3</sup>Univ. Claude Bernard, Lyon, <sup>4</sup>JGI, Walnut Creek, CA



## INTRODUCTION

*Frankia* are nitrogen-fixing actinomycetes, high G+C gram-positive actinobacteria that form root nodules on ecologically important actinorhizal plants. Actinorhizal plants occur in seven families that are only distantly related to each other. Symbiotic interactions between *Frankia* and the host plant are still not well understood. Genomes of three *Frankia* strains representing extremes of genome size and host range were compared to try to identify genes that are involved in host specificity, root infection, and lineage specific gene family expansions. The genome of *Frankia* CcI3, a strain that infects only *Casuarina* sp., and *Frankia* strain EanIpec, a broad host-range strain that infects plants from three families, were sequenced in collaboration with the Joint Genome Institute, while the third genome, *Frankia* strain ACN14a is being sequenced by Genoscope, France. Note that all three genome project are still at an unfinished stage. Analyses were done using the data available as of March 2005.

Preliminary analyses indicate that the difference in genome size is mainly due to lineage specific gene amplification in EanIpec and ACN14a rather than gene losses in Cci3. The larger genome size is not due to a higher percentage of poorly characterized genes and ORFans

## GENERAL PROPERTIES OF THE GENOMES

All three *Frankia* genomes have a G+C content above 70% (**Table 1**). The length of the genomes varies between 5.4Mb for the smallest genome (CcI3) to 9.1Mb for EanIpec. The genome size is reflected in the number of predicted protein encoding genes. Which processes were mainly responsible for the change in genome size: gene loss, gene acquisition by gene transfer, or lineage specific gene duplications?

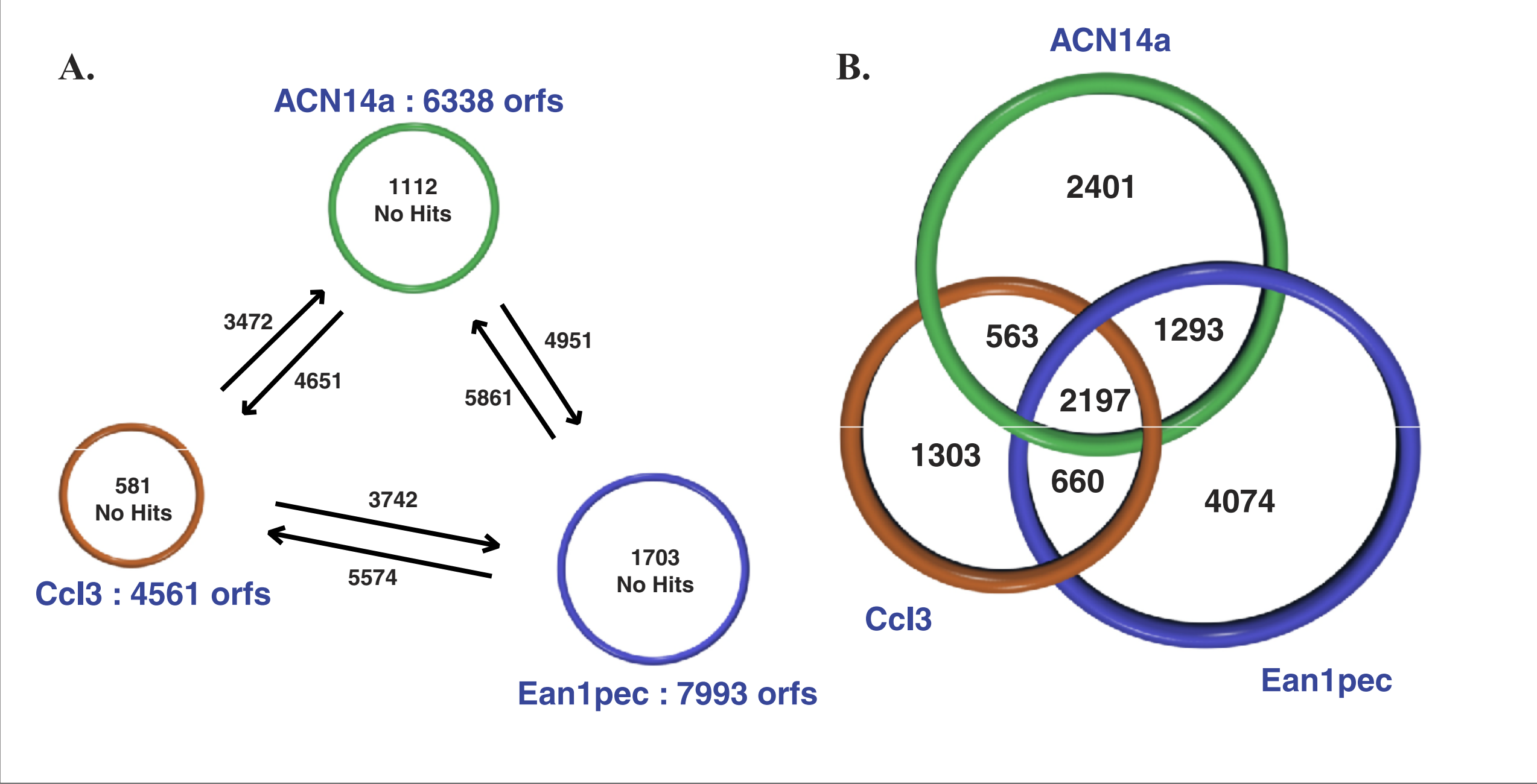
Classification of genes according to function reveals that gene numbers in some functional categories do not always follow the general trend reflected in genome size. In some categories (Cell division and chromosome partitioning; Nucleotide transport and metabolism; Translation, ribosomal structure and biogenesis) the absolute numbers stays constant. Only two subcategories (lipid metabolism; secondary metabolite biosynthesis, transport and metabolism) increase much more than the genome size.

	Frankia CcI3	Frankia ACN14a	Frankia EanIpec
Chromosome length (nt)	5.4 mb	7.5 mb	9.1 mb
GC%	70.10%	70.81%	70.90%
Total candidate protein-encoding gene	4561	6338	7993
tRNAs	46	--	49
<b>Information storage and processing</b>			
J Translation, ribosomal structure and biogenesis	164(3.6%)	170(2.7%)	169(2.1%)
A RNA processing and modification	1(0.02%)	1(0.02%)	1(0.01%)
K Transcription	287(6.3%)	511(8.1%)	596(7.5%)
L DNA replication, recombination and repair	366(8.0%)	240(3.8%)	533(6.7%)
B Chromatin structure and dynamics	1(0.02%)	1(0.02%)	1(0.01%)
<b>Cellular processes</b>			
D Cell division and chromosome partitioning	40(0.9%)	39(0.6%)	39(0.5%)
Y Nuclear structure	0	0	0
V Defense mechanisms	50(1.1%)	82(1.3%)	86(1.1%)
T Signal transduction mechanisms	208(4.6%)	304(4.8%)	366(4.6%)
M Cell envelope biogenesis, outer membrane	181(4.0%)	211(3.3%)	248(3.1%)
N Cell motility and secretion	10(0.2%)	2(0.03%)	8(0.1%)
Z Cytoskeleton	0	3(0.05%)	0
W Extracellular structures	0	0	0
U Intracellular trafficking, secretion, and vesicular transport	38(0.8%)	30(0.5%)	48(0.6%)
O Posttranslational modification, protein turnover, chaperones	118(2.6%)	137(2.2%)	146(1.8%)
<b>Metabolism</b>			
C Energy production and conversion	232(5.1%)	396(6.3%)	442(5.5%)
G Carbohydrate transport and metabolism	191(4.2%)	261(4.1%)	347(4.3%)
E Amino acid transport and metabolism	281(6.2%)	445(7.0%)	492(6.2%)
F Nucleotide transport and metabolism	87(1.9%)	88(1.4%)	91(1.1%)
H Coenzyme metabolism	164(3.6%)	179(2.8%)	176(2.2%)
I Lipid metabolism	170(3.7%)	404(6.4%)	481(6.0%)
P Inorganic ion transport and metabolism	170(3.7%)	260(4.1%)	281(3.5%)
Q Secondary metabolites biosynthesis, transport and catabolism	169(3.7%)	346(5.5%)	464(5.8%)
<b>Poorly characterized</b>			
R General function prediction only	502(11.0%)	883(13.9%)	988(12.4%)
S Function unknown	210(4.6%)	271(4.3%)	309(3.9%)
No COG homologs found	1649(36.2%)	1868(29.5%)	2700(33.8%)

**Table 1.** General genome features and functional assignments according to BLAST searches performed against the COG Database using a cutoff E-Value at 10e-04.

## PROTEOME COMPARISONS

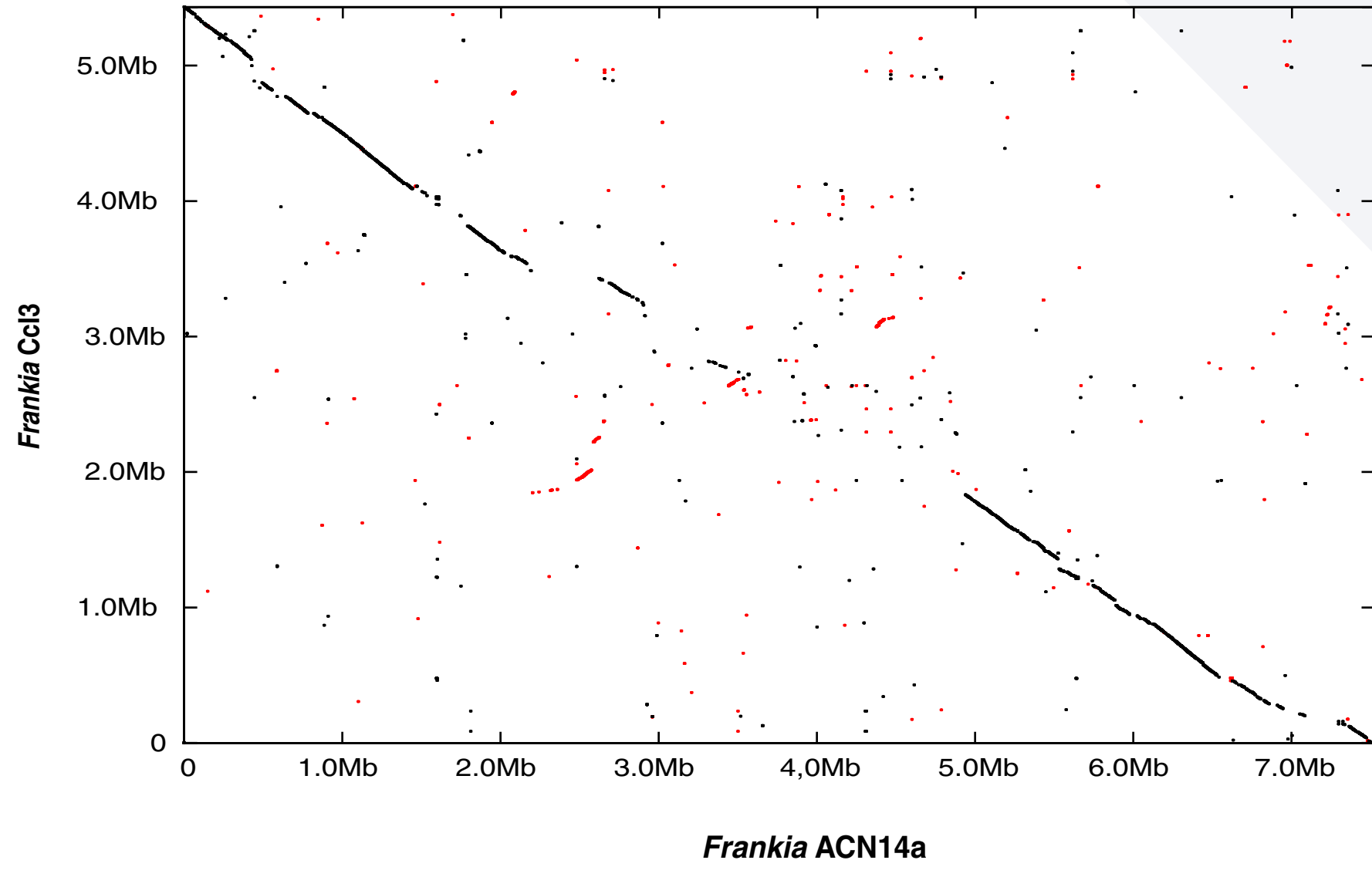
Results from Blast searches performed between the three genomes are summarized in **Figure 1**. 581 predicted proteins (12.7%) in the smallest genomes (CcI3), 1112 (17.5%) for ACN14a and 1703 (21.3%) for EanIpec do not have any recognizable homolog in the other *Frankia* genomes (**Fig. 1A**). Using a reciprocal blast hit scheme (Montague and Hutchison 2000; Zhaxybayeva and Gogarten 2002), we detected a total of 2197 orthologous sets of genes shared between the three genomes (**Fig. 1B**). The discrepancy between genes without orthologs (4072 in EanIpec) and the ones without homology (1703 in EanIpec) provides a first clue that many of the additional proteins in the larger genomes are due to gene family expansions. 67.4% of all ORFs present in the three *Frankia* genomes (counting all homologs separately, **Fig. 1A**) have a recognizable homolog in all the three strains. Only 11.6 % of the combined non-redundant set of genes is present in all three genomes (counting orthologs only once, **Fig. 1B**).



**Figure 1. A)** Non-reciprocal Blast hit and **B)** Reciprocal Blast hit results for protein-protein comparisons. Blasts were performed using an E-Value cutoff of 10e-04 with a word size of 2. 67.4 % of all genes (i.e. 12752 genes) belong to gene families that have a representative in each of the three genomes (A), whereas the core as defined by Welch et al. (Welch, Burland *et al.* 2002) consists of only 11.6% (B).

## DNA-DNA DOT PLOT

A genome-genome DNA sequence comparison between ACN14a and CcI3 reveals extensive synteny and only few rearrangement events between the two genomes (**Figure 2**). A strong diagonal is visible throughout the comparison. Only few insertion, deletion and inversion events were detected. The majority of these events are centered around the terminus of replication.



**Figure 2.** Nucleotide genome dot plot generated by MUMmer 3.0 (Kurtz, Phillippy *et al.* 2004) between ACN14a and CcI3. Dots or lines represents region of maximal matches between the 2 genomes over a minimum nucleotide window size of 30. The black indicates matches on the forward DNA strand and red the complementary DNA strand. The coordinates for comparison begin at DnaA, which is close to or at the origin of replication.

## GENE FAMILIES

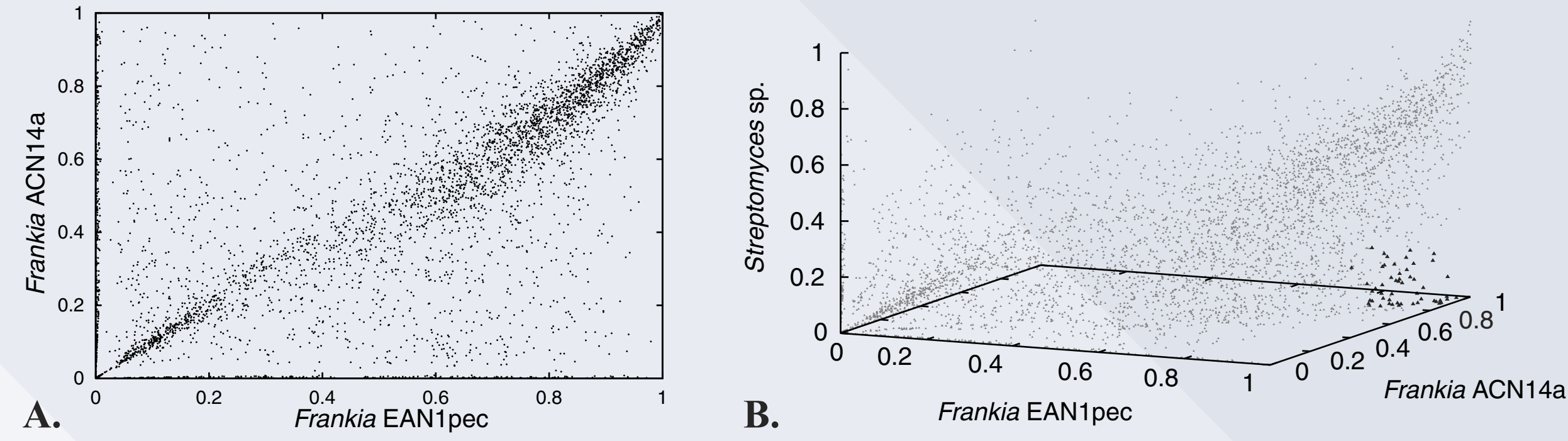
Looking at each gene family, there is clear evidence for ACN14a and EanIpec having tremendously expanded gene families (**Table 2**). For example, there are 5 to 6 times more dehydrogenase genes in ACN14a and EanIpec as compared to CcI3. Acquisition of variant genes via HGT, gene duplication or loss might have contributed to the variation in gene family size. These gene family expansions might be a source for novel properties such as extended host range, symbiotic capabilities, and metabolic capabilities (see '**Genes potentially involved in symbiosis**' section for an example).

CcI3	ACN14a	EanIpec	Total	Predicted function
20	101	131	252	Dehydrogenase
42	100	106	248	Putative ABC transporter ATP-binding protein
30	64	75	169	WD-40 repeat protein
20	47	41	108	FadD8
17	36	48	101	Putative membrane transport protein.
8	41	43	92	Putative acyl-CoA dehydrogenase
12	25	52	89	CYTOCHROME P450
12	21	45	78	Putative two-component system response-regulator
4	35	34	73	Putative enoyl-CoA hydratase
11	23	38	72	Multi-domain Polyketide synthases
13	25	24	62	Hypothetical protein
6	22	31	59	Putative Betaine Aldehyde Dehydrogenase (BADH)
2	23	33	58	Putative fatty acid-CoA racemase
11	15	29	55	Sensory box protein
...	...	...	...	...

**Table 2.** Subset of gene families presents in all three genomes. Gene families were detected using Tribe-MCL, an algorithm using a Markov Clustering method for large-scale detection of genes families (Enright, Van Dongen *et al.* 2002).

## BLAST SCORE RATIO (BSR) PLOTS

To assess the level of protein conservation, for each ORF in CcI3 we calculated the bitscore ratio for the best hits on ACN14a and EanIpec, respectively, over the bitscore for the comparison against itself (**Figure 3A**) (modified from Read, Myers *et al.* 2003). Genes mapping onto the lower left hand corner represent poorly conserved genes; genes in the upper right hand corner correspond to genes conserved in all three genomes, whereas dots falling off the diagonal represent genes conserved in CcI3 and only one of the other *Frankia* genomes (**3A**). Adding the bitscore ratio for two *Streptomyces* species (best hits of either *S. avermitilis* or *S. coelicolor*) reveals genes highly conserved in all three *Frankia* genomes but absent or poorly conserved in Streptomyces. These are candidates for genes specific to symbiosis (Figure **3B**, dots in bold). They are mainly poorly characterized genes or with no COG homologues (33 genes), implicated in metabolic (14 genes, including *nifE*) and cellular processes (8 genes).



**Figure 3.** Blast Score Ratio plot (Bitscore ratios) obtained for each predicted ORF using CcI3 as reference genomes and compared with ACN14a + EanIpec(**A**) and *Streptomyces* sp.(**B**). Ratios close to 1 represent highly conserved protein. Bold dots in panel **B** are proteins with a ratio above 0.8 for both *Frankia* and lower than 0.2 in *Streptomyces* sp.

## HGTS, ORFANS AND GENE DUPLICATIONS

By comparing Blasts hits between all three strains and the NR database, 337 ORFans (ORFs with no hits in NR or in the other *Frankia*) were detected in CcI3 (7.4%), 752 in ACN14a (11.9%) and 1148 in EanIpec (14.4%). 834 ORFs potentially acquired through HGT were detected in CcI3 (18.2%), 1482 in ACN14a (23.4%) and 2459 in EanIpec (30.7%) (Potential HGT were identified as genes whose highest scoring significant hit in a BLAST search on NR scored higher than the hit(s) in the other *Frankia*). 10.4% of the ORFs in CcI3 (17.2% in ACN14a and 28.4% in EanIpec) had their highest scoring BLAST hit within the same genome, i.e., these genes probably represent lineage specific gene duplications. Lineage specific increase in gene family size is one of the main factors for the larger genome size (compare **Table 2**).

## GENES POTENTIALLY INVOLVED IN SYMBIOSIS

A number of processes are associated with the initiation and functioning of the symbiosis. Nodulation genes (*nod*) are known to be involved in signaling and initiation of root nodule development in rhizobial species; genes for nitrogen fixation (*nif* genes) are involved in the synthesis and functioning of nitrogenase and electron transfer to nitrogenase; hydrolytic enzymes may aid in penetrating host cell walls or otherwise digesting plant cell components during infection.

Presently, we have not found strong homology with any of the *nod* genes from *Rhizobium* although we cannot rule out the possibility that highly diverged versions of such genes may yet be identified. This suggests that the *Frankia* symbiosis differs fundamentally from the rhizobial symbioses in the initial steps of nodule induction. All three strains have highly conserved *nif* genes responsible for nitrogen fixation including orthologs of *nifV*, *nifHDK*, *nifENXWWZB*, *hesB/nifU*, *NifS* and putative genes for ferredoxin oxidoreductase and ferredoxin presumably involved in electron transfer. *nifV* is not contiguous with the other *nif* genes in EanIpec, suggesting that a rearrangement has taken place.

The annotated genomes were scanned for the hydrolases. Surprisingly, CcI3 has no obvious cellulase (endoglucanase) genes, contrary to previous reports of cellulase activity in this strain. EanIpec is the only strain with a strong hit for a pectate lyase gene. Three contiguous alpha-amylase genes are conserved among the three strains, and are homologous to those in *Streptomyces*.

<b>Table 3.</b> Summary of hydrolytic Enzyme Genes in <i>Frankia</i>				
	CcI3	ACN14a	EanIpec	
Cellulase	0	2	2	endoglucanase
Pectinase	0	0	1	pectate lyase
Amylase	3	3	3	alpha-amylase
Lipase	3	4	4	secreted lipase
Esterase	3	7	16	esterase/lipase

## ACKNOWLEDGEMENTS

Support provided by the NSF/USDA Interagency Microbial Genome Sequencing Program and the DOE. We thank Genoscope for access to the ACN14a genome sequence.

## REFERENCES

Enright, A. J., S. Van Dongen, *et al.* (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res 30(7): 1575-84.  
Kurtz, S., A. Phillippy, *et al.* (2004). "Versatile and open software for comparing large genomes." Genome Biol 5(2): R12.  
Montague, M. G. and C. A. Hutchison, 3rd (2000). "Gene content phylogeny of herpesviruses." Proc Natl Acad Sci U S A 97(10): 5334-9.  
Read, T. D., G. S. Myers, *et al.* (2003). "Genome sequence of *Chlamydomophila caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae." Nucleic Acids Res 31(8): 2134-47.  
Welch, R. A., V. Burland, *et al.* (2002). "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*." Proc Natl Acad Sci U S A 99(26): 17020-4.  
Zhaxybayeva, O. and J. P. Gogarten (2002). "Bootstrap, Bayesian probability and maximum likelihood mapping: Exploring new tools for comparative genome analyses." BMC Genomics 3: 4.